Automatic high fidelity foot contact location and timing for elite sprinting

Murray Evans¹ · Steffi Colyer² · Aki Salo^{2,3} · Darren Cosker¹

Received: 21 July 2020 / Revised: 9 July 2021 / Accepted: 5 August 2021 / Published online: 28 August 2021
© The Author(s) 2021

Abstract

Making accurate measurements of human body motions using only passive, non-interfering sensors such as video is a difficult task with a wide range of applications throughout biomechanics, health, sports and entertainment. The rise of machine learning-based human pose estimation has allowed for impressive performance gains, but machine learning-based systems require large datasets which might not be practical for niche applications. As such, it may be necessary to adapt systems trained for more general-purpose goals, but this might require a sacrifice in accuracy when compared with systems specifically developed for the application. This paper proposes two approaches to measuring a sprinter's foot-ground contact locations and timing (step length and step frequency), a task which requires high accuracy. The first approach is a learning-free system based on occupancy maps. The second approach is a multi-camera 3D fusion of a state-of-the-art machine learning-based human pose estimation model. Both systems use the same underlying multi-camera system. The experiments show the learning-free computer vision algorithm to provide foot timing to better than 1 frame at 180 fps, and step length accurate to 7 mm, while the system based on pose estimation achieves timing better than 1.5 frames at 180 fps, and step length estimates accurate to 20 mm.

Keywords Multi-camera · Computer vision · System · Sports · Measurement · Sprinting

1 Introduction

Making robust and accurate measurements of moving humans is a difficult task with a broad range of applications. In the entertainment sector, actors' performances can be used to drive animated characters [1,35] in film and video games, while measurements of human motion can be used throughout health and sports to monitor patients recovering from injuries [12,45], identify injury risks [18] and uncover determinants of sports performance [5]. Often these measurements are collected in dedicated studios and laboratories, but there is very broad scope for expanding to general environments and using purely passive sensors that avoid any interference with

the natural movements of the subjects. As well as the already mentioned applications in entertainment, injury prevention and sports coaching, this can further enable gathering of general sports analytics such as for football or tennis [21,55]. Machine learning has started enabling markerless pose estimation “in the wild”, but it is not yet clear that these general purpose systems can provide the accuracy required for the specific interests of health, sports and biomechanics [43]. More accuracy might be possible with specialised systems, but the accurately annotated datasets for niche measurements do not exist and may never exist. This increases the importance of understanding the accuracy of existing “off the shelf” general purpose systems in specific contexts.

This paper considers the specific task of localising and timing the foot-ground-contact events of an athlete during sprinting, enabling accurate measurement of step length and step timing, two major determinants of velocity. Longitudinal measurement of these determinants across a training season such as performed by Bezodis [3] can show an athlete's reliance on step frequency and directly inform training programmes, but the manual annotation performed in [3] is unrealistic for regular use by coaches. The gold-standard

✉ Murray Evans
m.evans@bath.ac.uk

¹ CAMERA, Department of Computer Science, University of Bath, Bath, UK

² CAMERA, Department of Health, University of Bath, Bath, UK

³ Present Address: KIHU Research Institute for Olympic Sports, Jyväskylä, Finland

technique for making these measurements is to use force plates embedded in a running track [19] and marker-based motion capture systems [46]. Although accurate, both these approaches have limitations that could be overcome by the use of a purely video-based solution. Force plates are expensive and few facilities can accommodate more than a couple of steps, while optical motion capture requires the athlete to wear markers which take time to emplace and can affect the athlete's natural performance.

Two video-based solutions for the measurement task are proposed and contrasted in this paper. The first is designed to directly observe foot-ground-contact events and improves upon the work of [15] with a new objective function for foot-location refinement. For the remainder of this paper, this approach will be referred to as “Occupancy-Based Step Measurement fOr Sprinting” (Obsmos). The second approach uses the general-purpose human pose estimation system OpenPose [10] as a base; with a novel multi-camera fusion and reconstruction approach to recover 3D pose tracks from which foot-ground-contact events can be inferred. This second approach will be referred to as Fused OpenPose.

The contributions of this paper are:

1. the Obsmos system that builds from the earlier work of [15] with a new objective function for foot-ground contact location refinement that is both simpler and more robust
2. the Fused OpenPose approach, which is the first 3D analysis of detecting foot-contact events and step lengths using a modern deep-learning-based general purpose human pose estimator.
3. the first comparative evaluation of two markerless foot-ground-contact analysis systems on elite sprinting datasets with ground truths.

The performance of the two presented approaches is compared on two datasets, both of which use a multi-camera system with the same general design. The first dataset was recorded using 5 cameras in a laboratory environment and allows for a ground truth using optical motion capture and embedded force plates. The second dataset is more real world and uses 9 cameras recorded at an indoor sprint training track. Ground truth for the second dataset is provided by manual annotation and an auxiliary high frame rate camera for timing. The first dataset had 18 runners, with a mix of trained and recreational runners, while the second dataset consisted of 14 trained athletes.

The remainder of this paper will discuss some related work (Sect. 2) and then describe the camera system used for acquisition as well as the data used for evaluation (Sect. 3). Next, the algorithms will be described, starting with the Obsmos approach (Sect. 4) and then the Fused OpenPose approach (Sect. 5). Finally, an evaluation of the algorithms versus the

ground truth will be presented, as well as a discussion of the results.

2 Related work

The progress of human pose estimation from images is well documented [11], starting with the manual annotation of sequences recorded to film camera and steadily adopting more automation as technology advances. Accurate tracking without manual annotation became possible through the use of optical motion capture. By palpating the body to locate specific bones beneath the surface, infra-red (IR) reflective markers can be placed upon the skin which have a well-judged position relative to the skeleton. By combining IR-sensitive cameras and lighting, the markers can be reliably tracked and positioned to millimetre accuracy [51]. A kinematic skeleton model can then be fit to the tracks of these markers giving estimates of the locations and orientations of bones and joints as the subject moves [42,44].

The accuracy of marker-based techniques is offset against the requirement to affix markers to the subject and the need for specialist cameras and lighting. In practice this reduces the adoption of these systems outside of laboratory conditions. The hope of robust video-based human pose detection algorithms is to facilitate measurement “in the wild” and without markers. The progress of the field is discussed in several reviews [11,25,53], and there is a clear general trend for systems to be based off of machine learning. The result are algorithms that provide detections of human body parts [48,49] of multiple people [10,20] from arbitrary camera views “in the wild”, which can even deduce 3D pose from single viewpoints [30,32].

These markerless human pose systems are based on large datasets [27,34] which consist of a set of images taken from the internet and manually annotated with the approximate locations of various joints for each of the people visible. The images cover a wide range of environments, scales and poses, including highly acrobatic poses. In general, the body part annotations in these training datasets have been made by minimally trained volunteers with no access to the subjects in the image, and not expert biomechanists, so there will always be some doubt about the accuracy of resulting models. However, creating training datasets that are both large enough to be robust and specific to a measurement task (e.g. training a network to specifically identify foot-ground contact from video) can be prohibitive, making it important to understand the applicability of general-purpose systems.

Several video-based solutions to the problem of foot-ground contact and step length measurement have been published previously. Zhu [57] uses motion blur to identify when a foot is in motion or static; however, the approach is likely to be very limited, especially as faster cameras are used

which are better at freezing fast motion in an image. Harle's work [23] uses a single camera and background subtraction. By adding together foreground masks (accumulating them) the feet can be identified when on the ground as peaks in the accumulator. Intuitively quite simple, a lot of processing is required to handle false detections which can more easily be removed with multiple viewpoints, and using a single camera, although convenient, will limit positional accuracy, especially if a large area is to be covered. The system of Dunn [13] is multicamera and may be the most comparable to the Obsmos system, but technical details are limited. They do provide a benchmark against which to measure performance however, reporting -4.9 ± 177.7 mm and 0.0 ± 0.03 s (Bland and Altman 95% limits of agreement) against manual digitisation. Rather than directly detect image features corresponding to contact events, there are approaches such as [29] which try to infer these events by tracking motion of the body over time. The evolution of this would be to use modern pose detectors alongside gait analysis techniques, such as the proposed Fused OpenPose approach.

Gait analysis [24,40] is the analysis of animal and human locomotion (walking, running etc.). Gait is often measured using a combination of force sensors (in the floor or in shoes), accelerometers [8,26] or marker-based optical motion capture. Specifically, the path of a marker over time [22,47], or the changing angle of certain body joints (knee, ankle, etc) [41], produces distinctive curves, sometimes referred to as gait curves (though the term is also applied to the changes in force/pressure of the subject's foot with the ground) which can be used to identify contact events. In [22] foot-ground-contact events are detected by identifying acceleration or jerk peaks on heel or toe markers of human runners over a range of speeds. For maximal sprints [38] observes that the rotation of the foot during take-off induces a distinctive dip in the height of the marker which can be reliably used to determine toe-off, while landing is best identified using acceleration peaks. In a related domain [47] compares approaches to identifying contact events in hoofed animals, including vertical velocity, acceleration and a trigonometric breakover feature similar to the toe-off feature noted by Nagahara [38]. When used on marker-based motion capture data with relatively little noise these methods produce good results and can inspire approaches for markerless motion capture such as the proposed Fused OpenPose system.

Markerless systems can capture human motion in a number of ways. Depth cameras such as the Microsoft Kinect are closely associated with algorithms that produce skeleton representations of human pose. Gait analysis from these skeletons has been investigated [2,52], but the sensors themselves have limited frame rates (typically 30 Hz) and use active illumination that can struggle in general environments, constraints which do not apply to normal cameras. Markerless systems for general cameras could fit 3D body

models to silhouette images, with later examples of this using person specific body models [14], but good results often required some manual intervention, laboratory conditions, and the solvers were extremely slow. More recently, human pose detectors such as OpenPose [10] have enabled relatively fast tracking of the human body in images. Some walking gait analysis has been performed using such detectors [39,54], but there is still concern about whether they can provide the kind of measurements biomechanists desire [43]. Concerns include the noise characteristics, precision of the detected features and extent to which the detections correspond to actual body parts. These factors can all be expected to impact what features can be observed in gait curves and how accurately gait events can be detected. The Fused OpenPose system presented in this paper takes inspiration from marker-based gait analysis and identifies what features can be robustly detected to identify ground contact events in sprinting.

The aim of this paper is to evaluate the performance of video-based sprint analysis systems based on a traditional computer vision approach of hand-designed features and on modern machine learning pose estimation. A number of machine learning-based pose-estimation algorithms could have been used for this experiment. Recent systems that report full 3D pose from monocular views such as VIBE [32] are impressive, but it is unclear how to best fuse the model positions between camera views and poor reconstruction of foot position can be seen even in video results presented by the authors. Systems which provide human body-part segmentations are also very powerful, but do not directly lead to trackable pose information—DensePose [20] remains the most detailed (with each pixel of the segmentation labelled for a specific point on the surface of the human body), but experience indicates the detection of feet is again poor. Of the systems that provide a sparse 2D detection of joints, OpenPose is (at time of writing) the only system to provide features on the foot and is easy to deploy.

3 Markerless sprint analysis camera system

Markerless detection of step length and step frequency can be done through a single-camera system over a limited section of track using systems such as [23], but single-camera systems have difficult error suppression and limited accuracy on 3D positional measurements. A multiple camera system can be designed to cover an extended area of track without resolution trade-offs and ensure accurate calibrated measurements.

The camera system presented in this paper follows the design proposed in [15]. This is a multi-camera system as shown in Fig. 1 that allows for an extended section of track and thus multiple steps to be observed. This range will allow coaches to capture information about different stages of the

athlete's run, whether the acceleration stage or peak velocity stages, etc. Cameras are positioned such that their views overlap and allow each foot-ground contact to be observed by at least three track-perpendicular cameras as well as the two track-parallel cameras. In this way, plenty of information is available for accurate 3D localisation as well as maximising the visibility of the exact moment of foot-ground contact. The height of the cameras is set at about waist height—a very low camera might be expected to make the foot-ground contact timing more obvious, but keeping the cameras at a neutral height means that full body motion capture should not be compromised by odd projection angles. So long as the resolution of the athlete in the images is not significantly affected, it is not expected that minor changes in the relative positions of cameras would affect performance of the system. More important is to ensure sufficient overlap of view and quality of calibration.

The camera system is expected to be synchronised. Ideally this would be through a hardware timing signal such as is commonly found on machine vision camera systems; however, the datasets presented later in this paper produced good results using broadcast TV cameras which could only be visually synchronised by external timing lights—in which case it is recommended that synchronisation be checked for each trial. More details on the used synchronisation are given in Sect. 6.

Different lengths of observation area can be configured using different numbers of cameras. For example, this paper presents two datasets, the first covering a length of approximately 8 m and the second covering approximately 16 m, using 5 or 9 cameras, respectively.

The camera system was calibrated using standard techniques. A circle-grid calibration board is shown to each camera in turn for intrinsic calibration [56] and then walked through the scene, ensuring it is seen by multiple cameras at any one time. Camera extrinsic parameters are calculated from these shared observations using Bundle Adjustment [50] to reach a globally optimal calibration. Calibration is set such that $z = 0$ is the floor plane, with $+z$ up. The y -axis was aligned to be parallel with the running track, such that runners run towards $+y$, to simplify processing.

4 Occupancy-based step measurement for sprinting

The first of the two approaches to foot-ground-contact detection that will be proposed in this paper builds off the design of [15] and consists of multiple processing stages. Foot contact events are detected, localised and approximately timed using a multi-camera processing algorithm, and then timing is refined using multiple instances of a single-camera algorithm. The overall system can be seen in Fig. 2.

4.1 Detection

The detection process first segments the athlete from the background, then creates occupancy maps which identify the presence of the athlete at locations in 3D space. The occupancy maps are examined over time to determine scene activity which provides approximate timing and location information for each foot-ground contact.

Given that the camera and background are static and only the athlete moves through the scene, a standard approach such as background subtraction (BGS) is effective at segmenting the athlete from the background. There are many approaches [7], but in general, they all work by a similar principle. Each pixel learns a model of the colour and brightness of the scene background, where the model will have allowances for sensor noise and colour variations caused by periodic non-salient motion of background objects (e.g. swinging tree branches). When an object moves across the scene, it will cause the colour of a pixel to change. Where that change is sufficient to no longer fit the background model, that pixel is labelled as “foreground”. The result is a set of foreground masks for each camera where, ideally, all background pixels are set to black, and only pixels inside moving objects of interest are set to white. In practice, BGS algorithms are imperfect and will struggle when there is a lack of contrast between foreground and background which can result in holes in the segmentation, or a need to accept more pixels being falsely labelled as foreground so as not to miss important body parts. For the foot-ground-contact application, it was found to be important to tune the BGS algorithm to get a full segmentation of the athlete's foot and not be so worried about the performance of body segmentation. It was also important that the athlete's shadow should not be detected as foreground. Reflections and background noise were mostly well handled by the multi-camera processing stage.

In extreme lighting conditions, where, for example, clouds cause rapid lighting changes, bright sunshine causes harsh, difficult to remove shadows, or where the background cannot be guaranteed to remain static, background subtraction could be replaced with a machine learning-based human-segmentation algorithm such as [20,33]. This choice does however come with a significant cost in speed (BGS algorithms can run at upwards of 20 fps, whereas the machine learning systems often run as low as 1 fps) and adjustability (tuning a BGS algorithm is well understood, the machine learning algorithms might not have the option to be tuned—though hopefully, the change to an ML system negates the need for tuning).

This paper stays with background subtraction, specifically, the IMBS-MT algorithm [6], which was chosen as it met the requirements for speed, shadow handling, intuitive tuning for the environment and an excellent reference imple-

Fig. 1 Camera system. Athletes run along the track (arrow) between a corridor of cameras. The diagram on the left shows a layout for 9 cameras, while the image on the right shows the setup as viewed by one of the track-parallel cameras. Cameras **a** look across the track (track-

perpendicular), while cameras **b** look along the track (track-parallel). The corridor can be extended to desired length by adding cameras, maintaining the 3-camera pattern (darker shaded)

Fig. 2 Overall system structure. Video from multiple cameras is input into the multi-camera processing stage. Here, the athlete is segmented from the image using background subtraction, and then further process-

ing identifies the location and approximate timing of each foot-ground contact. The resulting foot contact events are passed to a per-camera processing stage to refine the timing estimates

mentation. Examples of the recovered foreground masks, cropped to the region of the athlete, can be seen in Fig. 3. The biggest problems encountered with this approach were shoes or clothing with poor contrast against the floor or background.

The foreground segmentations from each camera view are fused together to create a set of “occupancy maps” by projecting them onto various horizontal scene planes: a ground plane, a knee plane and a body plane. Let the observable area of the scene be defined by an axis-aligned rectangle drawn on the ground. This rectangle can then be sub-divided into a grid of cells G of a given size c_s (Fig. 4) where the centre $\hat{p}_{rc}$ of each cell, $G(r, c)$ is a point in 3D space $\hat{p}_{rc} = [x_{rc}, y_{rc}, h, 1]^T$, and $z = h$ is the scene plane at height h . Each point $\hat{p}_{rc}$ can be projected into the foreground masks using the known camera calibration (Eq. 1) (where P_i is the projection for camera i):

$$p_{irc} = P_i(\hat{p}_{rc}) \quad (1)$$

The occupancy for a grid cell is computed as per Eq. 2:

$$O_G(r, c) = \frac{\sum_{i \in \mathbf{V}_{rc}} F_i(p_{irc})}{|\mathbf{V}_{rc}|} \quad (2)$$

In Eq. 2, $|\mathbf{V}_{rc}|$ is the number of cameras in $\mathbf{V}_{rc}$, and $\mathbf{V}_{rc}$ is the set of cameras in which p_{rc} is visible. Furthermore,

$F_i(x) = 1$ if the pixel $x = [u, v]^T$ is foreground, otherwise $F_i(x) = 0$.

The result is that the occupancy map (or *synergy map* [31]) is in effect an image showing the additive projection of each camera’s foreground image to the specified scene plane (Fig. 5). The more cameras that see a cell as foreground, the brighter that pixel is in the occupancy map. Significantly, when a sprinter runs through the scene, cells in the ground occupancy map will be at full brightness only when the runner’s foot is on the ground.

Ground plane occupancy is used for detecting the athlete’s feet on the ground, knee-plane occupancy is used for managing problems that can occur when the athlete’s non-contact foot causes occlusions of the on-ground foot, and body occupancy is used for preventing false detections [15]. A cell is considered to be occupied or *active* in a given frame if cell occupancy passes a pre-defined threshold. The threshold is a tunable parameter that indicates what proportion of the cameras are required to show the cell as foreground, allowing for any partial detections within the image region corresponding to the grid cell. If the feet are likely to be only partially detected, or if the foreground masks are particularly noisy, a lower threshold might be needed, but setting it too low will allow for false detections, particularly if it is set low enough that it can be activated by only one camera.

A temporal processing step is then used to detect and localise foot-ground-contact events. The thresholded occupancy map identifies which cells are active on a given frame.

Fig. 3 Background subtraction examples for a typical frame of video. Segmentation failures that cause noise in the background or “holes” on the body were deemed to be acceptable so long as foot segmentation was reliable and shadows on the floor were well handled

Fig. 4 The region of interest of the track can be divided into a grid (actual grid cells smaller than shown in this image). Foot-ground contacts can be detected by checking for occupancy of the grid cell at each time instant

Fig. 5 Given the foreground masks of Fig. 3, three occupancy maps are created. From left to right, these are the ground plane occupancy, knee plane occupancy and body occupancy. When each map cell is projected to a camera, it projects to a region of the foreground mask that is labelled as either foreground or background. Cells which are brighter in the occupancy map project to foreground regions in many cameras, while dark cells are foreground in none. Cells in the ground plane occupancy will be brightest when the foot is present and on the ground at

their 3D location. Equally, knee-plane cells will be brightest when the leg is intersecting the knee-plane, and body cells will be brightest when the body is passing through that region of space. The ground plane and knee plane occupancies are shown before thresholding; where colour is available, the green region on the body occupancy shows the region that passes the threshold value, indicating those cells are *active* for this frame. Note that the foot contact can be clearly observed despite the problems (holes, noise) with the foreground masks

Activation periods for each cell are recorded in an *activity map*, where an activation period is defined by the frame a cell becomes active to the frame before a cell becomes inactive.

Ground contacts are detected by processing the activity maps. A contact is observed as a cluster of activations in space and time. Thus, time is played forwards frame by frame:

1. Each cell is checked to determine whether the current frame is within an activation period.
2. The earliest start f_s and latest end f_e of all current activation periods are determined.
3. A cell is considered to be active in the current frame if it has an activation that starts before f_e and ends after f_s .
4. A contact is created in the current frame from a contiguous cluster of currently active cells.
5. Contacts in the current frame are compared with existing contacts from previous frames and merged if they share the same location in space, thus producing temporal consistency.
6. The duration of a contact, its start and its end, can be determined from the activation periods of the cells.
7. A contact is only deemed valid if it overlaps with the body occupancy for the duration of its existence.

4.2 Foot-ground-contact position refinement

Here, an updated version of the algorithm from [15] is proposed. The general approach remains the same: A foot-sized bounding box is initialised at a point in 3D on the ground plane and then has its 2D position on the ground plane optimised to minimise an objective function. The proposed update changes the features used in the objective function to both simplify and improve robustness.

Firstly, an image is produced for each camera which accumulates the foreground masks for every frame between the contact start frame f_0 and end frame f_k , as per Eq. 3.

$$A_i(x) = \frac{1}{k} \sum_{j=f_0}^{f_k} F_{ij}(x) \quad (3)$$

In Eq. 3, $F_{ij}(x) = 1$ if the pixel $x = [u, v]^T$ is labelled as foreground in the mask image from camera i on frame j , otherwise $F_{ij}(x) = 0$.

A threshold τ_A is applied to the accumulator images to create a binary mask α_i with pixels larger than the threshold set to white. Next, a distance transform [16] is calculated to produce D_i , an image where every pixel has a value based on its distance to the nearest white pixel in the binary mask. An example of these three images (the accumulator, the result of the threshold and the distance image) can be seen in Fig. 6.

Secondly, an image of the ground occupancy for the duration of the foot contact is created. This is simply an image

with white pixels where occupancy map cells were active during the contact and black otherwise. A distance transform of this map is then computed giving the image D_s .

A bounding box can be defined with length b_l , width b_w and height b_h and oriented such that the length axis is parallel to the running direction (if the calibrated camera system is configured as in Sect. 3 then this can be an axis-aligned bounding box with length on y , width on x and height on z).

The bounding box can have the approximate size of a human foot and can be centred on the ground plane at $\hat{p}_b = [x, y, 0]^T$. The eight corners of the bounding box can be projected into each camera view and the smallest encapsulating bounding boxes determined.

$$\epsilon(x, y) = \epsilon_S(x, y) + \sum_{i=0}^n (\epsilon_{D_i}(x, y) + \epsilon_{C_i}(x, y))$$

The first term $\epsilon_S(x, y)$ in this error is simply the value of $D_s(r, c)$, where (r, c) are the occupancy map cell coordinates corresponding to a scene position $\hat{p}_b$. This ensures that the minimisation search cannot move too far away from the initially estimated contact location. Without this, the error can suffer from problems when the projected bounding box gets too close to, or leaves, the bounds of the camera views. $\epsilon_{D_i}(x, y)$ is the sum of values inside the bounding box within the distance image D_i , normalised by the area of the bounding box. $\epsilon_{C_i}(x, y)$ is a term designed to help the bounding box to centre on the foot. The distance δ_l between the left edge of the bounding box and the left-most white pixel in α_i , and the equivalent distance on the right δ_r are determined (Fig. 7). With those, the term is $\epsilon_{C_i} = |\delta_r - \delta_l|$.

The error can be minimised using any standard derivative-free search algorithm. Results presented in this paper use the SBPLX algorithm from the NLOpt C++ library [28].

Adjusting the values of τ_A and the size of the foot-sized bounding box does have an effect on the overall performance of the algorithm. During a contact event, the foot is not wholly static, and thus, different pixels will be classed as foreground for different proportions of the estimated contact time. Figure 9 shows the stages of a sprinter's typical foot-ground contact—landing on the toes, absorbing the impact, rolling forward and springing off the toes again. If τ_A is set very large, then only pixels that are static for the whole contact will pass the threshold. As τ_A is reduced, more pixels will pass the threshold. The aim when setting τ_A is to ensure that the whole foot is segmented without capturing pixels only active transiently as the foot enters or exits the contact location. After empirical analysis, the optimal value was found to be in the region of 0.25, much smaller than this and the foot would not be well-segmented, impacting position estimation. Generally, values of up to 0.5 could be used before

Fig. 6 Crop of the region around the person for, from left to right: accumulator of foreground masks for a contact, thresholded result and the resulting distance transform

Fig. 7 Diagram of the error term to aide centring the box on the foot. The optimisation will seek to make δ_l and δ_r equal

performance significantly dropped off, but 0.25 was optimal in experiments.

Given the optimisation error, it might be expected that the foot-box be sized large enough to fully enclose the foot, but not so large as to run the risk of capturing too much of the leg or other distractions. As each runner has a different size of foot, it might also be expected that the size of the box will benefit from tuning to each individual. However, upon testing, it was found that, while the optimal box size was different for each individual, these optimal sizes did not correlate with the physical size of the runner's feet. This is likely to be at least partly caused by the different poses runner's feet can take—some out-turned, some never flat on the floor—making the optimal box size slightly unpredictable. However, the optimal box size for each runner was only marginally better than the best-on-average size. The best-on-average size was 400 mm long by 20 mm wide by 60 mm high, but sizes in the general range of (350, 20, 60) to (400, 60, 60) (length, height, width) performed similarly. As might be expected, this best-on-average box size is longer than any real foot allowing the box to always fully enclose the foot and for the centring term to neatly centre it. The height of the box is taller than a typical foot, but allows for sprinters who run on their toes (thus the foot does not go flat). The most surprising measurement is the width, which is substantially narrower than a real foot. However, when projected into the images, this narrow foot-box most often gives the neatest 2D bounding box around the foot. A comparison image of a wide and narrow foot box is shown in Fig. 8.

4.3 Foot-ground-contact timing refinement

The multi-camera process only gives approximate timing for the start and end of a foot-ground contact. To achieve a more accurate timing estimate a per-camera process is used which involves tracking simple image features as the foot makes contact with or departs from the ground. [15] observed that the most significant feature to determine contact start and end time was the vertical motion of the foot and thus took a small window of the image around the foot and divided it into vertical slices. Each vertical slice could then be easily tracked from frame to frame to identify when vertical motion starts or ends. Examples of a typical sprinter's foot-ground contact, and the slice features, are shown in Figs. 9 and 10, respectively.

As this process is performed per-camera, [15] also identified how to best combine the results of the multiple cameras that observe each foot contact, deducing that it was best to take the earliest reported landing and the latest reported take-off.

5 Foot-ground-contact detection from multi view pose estimation

Machine learning is an appealing approach to solving computer vision problems however such solutions require training with large datasets to achieve robustness against all the variations in a domain—this can include image variations

Fig. 8 Comparison of wide and narrow foot box. The smallest 2D bounding box encapsulating the 3D bounding box is a neater fit to the foot when using the narrower foot box

Fig. 9 Typical sprinter's contact

Fig. 10 Construction of slice images. The image window around the foot (left), the slice divisions, the signed image gradient and the colour and gradient slice features

caused by lighting, environment, weather and variations in the movement patterns, shapes and colours of the object or event to be detected. There are several ways of taking advantage of machine learning to solve the foot contact detection and timing problem for which the Obsmos system was developed.

Firstly, one could attempt to train a system specifically for detection of contact events. Such a system does not currently exist and so not only would it need to be designed, there would need to be a substantial data collection to create a training data that captured a wide enough variation of lighting conditions, runners, shoes, tracks. Creating this dataset is a prohibitive requirement.

An alternative approach might be to replace parts of the Obsmos system with machine learning alternatives. One obvious target for such a replacement would be the system-critical background subtraction stage. BGS almost always requires on-the-day tuning to get performance that handles the available lighting conditions and background appearances. In the worst case, these conditions can change over time, and it can be a constant battle to balance between

correctly handling shadows and having sufficient sensitivity to fully segment the object of interest. Powerful machine learning based systems exist [20,33] which can segment out humans in images and which can alleviate many of the problems of BGS, but the trade-off is significantly increased computation time (seconds per frame rather than tens of frames per second) and the precision of the segmentation can still struggle with body extremities such as feet which are critical in the foot-contact task.

A third option would be to take advantage of recent human-pose estimation systems and so mimic the approach of foot contact detection from motion capture [4,36,37]. One example [2] of such a system was based on using Microsoft's Kinect V2, but a foot contact detection should also be possible from systems such as OpenPose [10] which can be applied more generally to any camera system.

The remainder of this Sect. will describe an experiment that follows this third approach. OpenPose [10] will be used for human detection in the images, and an algorithm proposed to fuse those detections between camera views, and then a

novel algorithm proposed for estimating foot contact events from the fused 3D tracks.

5.1 Human detection in images

OpenPose [10] is one of many recent markerless full-body pose detection systems. Given an image of a person, a deep neural network estimates the location of various body parts. OpenPose is particularly appealing for foot contact detection because it detects points on the foot [9], it has been demonstrated to have good robustness in a wide range of environments and is relatively easy to make use of.

To determine step characteristics, OpenPose is applied to each camera view independently, yielding single-camera identifications of the people and their feet in the scene, as seen in Fig. 11.

During testing, it was observed that OpenPose made several common mistakes on the images. Firstly, it would mistake the tripods in the background of the image as people. It would also struggle to correctly identify the left versus the right sides of the running athlete. It could also be prone to losing one or both legs.

Much as with the Obsmos system, covering an extended length of running track will require the use of multiple cameras and so it makes sense to also take advantage of multiple cameras for rectifying detection errors. As a result the camera system used for the OpenPose-based approach is the same as for the Obsmos system. A system is proposed for fusing together the multiple camera views and producing a track of the person's joints through the monitored section of track. The proposed system is termed "Fused OpenPose" and will be fully described in the remainder of this section.

It should be noted that OpenPose provides a limited 3D reconstruction tool which is not used in Fused OpenPose. The OpenPose tool does not resolve the cross-camera matching problem and is designed for specific camera hardware and is not being maintained.

5.2 Fused OpenPose

The processing stages of the Fused OpenPose approach are shown in Fig. 12. OpenPose is first applied to each camera view, and then the fusion process combines the views together into a 3D skeleton. The joints of this skeleton are tracked through the full video sequence. Finally, the tracks of the toes are analysed to identify individual foot contact events. By fusing as early as possible (just after the per-camera application of OpenPose) the system can try to rectify the errors on OpenPose and consistently resolve passing between cameras.

At any moment in time, each camera will provide a single image, and OpenPose will report any detections in that image. Multiple people may be reported in one single image,

whether real or false, and so the first stage is to perform cross-camera matching. To associate person detections between views, inspiration is taken from the occupancy maps used by [15] and the Obsmos system. The volume of the area being run through is subdivided into a grid of cuboids. The grid is a single layer tall, with each non-overlapping cuboid being 0.25 m by 0.25 m by 2.0 m. Each cuboid can then be projected into the camera views resulting in a 2D-bounding box.

Each person in each view is tested against the projected bounding box to determine if their neck-point and mid-hip point are contained within the projected bounding box, and the cuboid scored by the number of viewpoints that have a person within the bounds of the projected cuboid. High population cuboids indicate the approximate 3D extents of persons detected in the scene and also group together the relevant detections. As only one runner is expected to be in the active area at any one time, it is simple to then take the highest population cuboid as the runner. Tripods and other false detections occur in only a subset of the cameras, or outside of the observation area, and as such do not generate an occupancy peak and so are quickly discarded.

Once the per-camera detections have been associated, the body parts of each person can be reconstructed in 3D. The 2D detection in each view can be back-projected to produce a ray in 3D space and the rays intersected to provide the reconstruction. To make this more robust, a RANSAC [17] approach is used. Pairs of rays are selected at random to compute the 3D point. The distance of this 3D point to each of the rays is computed and rays with distances smaller than a pre-specified threshold labelled as inliers. The largest set of inliers is then selected to produce the final intersection point. One of the main problems with OpenPose is a tendency to mis-label left and right body parts. For body parts that come in pairs, the RANSAC process is adapted to solve for two points. Consider an elbow: All elbow detections (left and right) are collected. First, any view that has two detections is selected. Detections from other views are randomly paired with one of those initial selections, thus forcing the process to produce two elbow reconstructions. The solution with the largest number of inliers for both points is taken. Left and right are then deduced using known scene geometry (the runner is known to be running parallel to the y-axis of the scene and towards +ve y).

Having the 3D skeleton in each frame results in a track of all the joints over time, meaning that there is a track of the OpenPose toes similar in principle to the track of a toe marker in marker-based motion capture. However, the Fused OpenPose track is substantially noisier, as can be seen in Fig. 13

When using marker-based motion capture, foot-ground-contact events can be discovered by identifying one or more features in the vertical track of the toe marker. These can

Fig. 11 OpenPose detects joints of the human body, shown here as circles. In these images, lines connect the ankles-knees-hips and wrists-elbows-shoulders. Black-on-green lines show left limbs, and white-on-red lines show right limbs. Green cubes and red cubes show the projection of the recovered 3D joints (green=left, red=right). Open-

Pose was found to be prone to a few errors, including being confused by tripods (left), swapping left-right limbs (centre) and merging left/right limbs together (right). By using multiple cameras and robust reconstruction, these errors can generally be handled

Fig. 12 Processing staged of the Fused OpenPose-based foot-ground contact event timing system

Fig. 13 Comparing the track of a motion capture toe marker versus a Fused OpenPose toe (temporal alignment approximated). Acceleration events for foot landing and take-off are obvious in the marker track but hidden by noise in the Fused OpenPose track

include the distinctive acceleration peaks on landing and take-off, the jerk of those acceleration peaks and a distinctive dip in the track of the toe-marker as the foot rotates ready for take-off [38]. These features are visible in Fig. 13, but are very unclear in the Fused OpenPose toe track. With good smoothing, such as from a Kalman smoother, much of the noise can be filtered off, but there can still be multiple acceleration peaks for each contact event, or the acceleration can

appear elongated making it less clear when contact actually starts or ends.

To get a robust estimate of contact times from the Fused OpenPose track, the following approach is used.

1. Smooth vertical motion of the toe using a Kalman smoother with a constant acceleration model.

Fig. 14 Effect of different levels of smoothing, showing the toe height over time (top), velocity (middle) and acceleration (bottom). Vertical lines show the approximate times of landing and take-off. Acceleration

peaks that correspond with landing and take-off can be seen with the correct level of smoothing, but are hard to identify without smoothing and totally corrupted with too much smoothing

2. Identify when the foot is on the ground using negative-to-positive velocity changes and height analysis.
3. Identify the most appropriate acceleration phase for the landing and take-off.
4. Find the peak acceleration for those phases.

5.2.1 Track smoothing

The vertical motion of the toe point is smoothed using a Kalman smoother. The process noise and measurement noise for this filter are determined through experimentation, but it is clear that the amount of smoothing can greatly affect robustness, accuracy and precision. Various smoothing levels can be seen in Fig. 14, showing how the acceleration peaks corresponding to landing and take-off can be difficult to distinguish from non-salient acceleration peaks, but can also be corrupted by excessive smoothing.

5.2.2 Contact detection

During running, the foot will come down towards the ground (thus have negative vertical velocity), then remain relatively stationary on the ground, then rise away from the ground (thus have positive vertical velocity). To detect when the foot is on the ground, negative-to-positive zero-crossings of the

toe track's vertical velocity are found. The resulting points can include moments where the foot follows an arcing path as it swings through from back to front, but these moments can be filtered out based on foot height. An example can be seen in Fig. 15.

5.2.3 Identifying landing and take-off frames

Landing and take-off are significant accelerations and should thus be identifiable in the acceleration of the toe track. Ideally there would be one distinct deceleration for the landing that occurs just before the detected zero-crossing and one distinct acceleration for the take-off that occurs just after the zero-crossing. In practice the acceleration peaks are not so clearly unique so a robust method is required to find the correct moment.

First, vertical velocity and acceleration are calculated from the height of the smoothed toe track. Next, the regions of positive acceleration are identified (positive acceleration means upwards acceleration) and identified by their start and end frames. For a given contact, the potential acceleration regions are identified as having a minimum height less than 40 mm above the height at the zero-crossing frame and be within 40 frames of the zero-crossing frame. Landing and take-off acceleration phases are differentiated by the sign of and mag-

Fig. 15 Negative-to-positive zero-crossings of the toe's vertical velocity can be used to identify when the foot is on the ground. Filtering is applied to ensure these points are at local minima of the toe height.

Vertical lines show the zero-crossings, with dotted lines being removed and the solid lines being retained by filtering

nitude of the velocity at the start and end of the phase. The best of these potential acceleration phases is selected as the one that has the largest change in velocity over the course of the phase. Finally, the landing or take-off frame is identified as the frame with the largest acceleration within the selected phase.

6 Performance evaluation

Two datasets were used to evaluate and compare the performance of the Obsmos and Fused OpenPose systems. In both datasets, Sony PXW-FS7 cameras were used to record the data at HD resolution and 180 fps. To synchronise the cameras and secondary systems (optical motion capture, force plates, high-speed camera), an operator would initiate a trigger pulse on each run. That trigger would initiate recording on ground-truth optical motion capture system and force plate, or high-speed camera. It would also trigger a set of timing lights. Each bar of lights consisted of 20 lights which would light in sequence over 20 ms. As every camera could see at least 1 bar of lights, the recordings could be temporally aligned after the fact by counting how many lights were illuminated and offsetting each camera's recording appropriately. The environments of the two datasets can be seen in Fig. 17.

Both the Obsmos and Fused OpenPose systems were run on a Linux-based workstation with an NVidia CUDA capable

graphics card. The Body25 model of OpenPose was used as it includes foot and toe features and was processed on the GPU. Neither system should be considered real-time, primarily because there are many cameras running at 180 Hz which produces a large number of images to process.

6.1 Dataset 1—Indoor laboratory sessions

The first dataset used a 5 camera setup covering a running corridor of approximately 8 metres as shown in Fig. 18. This is a laboratory-based environment with 2 force plates (Kistler, 9287BA) operating at 1000 Hz and an optical motion capture system (Qualisys, Oqus 400) consisting of 10 cameras operating at 250 Hz. The force plates were used for gathering ground truth timing information for one or two contacts on each run. The motion capture system was used to capture ground truth step length data on a subset of the runs. A mix of 8 trained sprinters and 10 recreational runners made 10 runs each through the space. For the 10 recreational runners, 5 of the runs were made wearing motion capture markers on the feet. Running speeds on this data range from 5 to 7 m/s for the trained athletes, while most of the recreational runners were nearer 2 to 3 m/s.

This was an indoor laboratory environment with minimal external influence. Lighting was artificial and constant across all recording. The lighting was intentionally diffused by bouncing off of the laboratory walls to minimise shadows that might confuse the background subtraction algorithm.

Fig. 16 Top: smoothed toe track showing time of detected zero-crossings (indicating foot on ground) and 40-frame search limits. Middle: toe velocity. Bottom: acceleration of toe track, with regions of positive acceleration highlighted. The final choice of landing and take-

off time is made as the peak acceleration for the acceleration region with the largest acceleration change within the search limit and the minimum height of the foot during the acceleration region

The lighting is, as a result, not very bright. To successfully freeze motion the cameras required a high shutter speed, with the result that the video data contains a significant quantity of sensor noise. The ground was a mid-grey, which proved to be less than ideal for many of the runner's shoes (particularly the sets with trainers—often dirty white and thus mid-grey), requiring a difficult balance to the BGS parameters which could not exclude all image noise.

This dataset provides accurate timing comparisons for one or two steps per run of all runners as they contact the force plates in the floor of the laboratory. Accurate step position information is provided on the runs where motion capture markers were worn.

Timing errors for the dataset, divided into recreational and trained runners, are shown in Table 1. Step length errors for the recreational runners when they were wearing markers are shown in Table 2.

6.2 Dataset 2—Indoor sprint training track

The second dataset used a 9 camera setup as shown in Fig. 1 allowing for a 16-metre corridor at an indoor running track using the available natural light. This dataset used only trained athletes simulating a sprint training session. There were 14 athletes (12 sprinters, two hurdlers) totalling 70 runs through the scene. The track environment prevented using force plates for verification, while markers were not used to

be as unobtrusive as possible. As such, ground truth data for step length estimates consisted of the manual annotation of the toes of each athlete on each visible step in all cameras. All annotations were made by a single person, and a subset of annotations was repeated to verify consistency. There were approximately 8 contacts per run. For step timing, a high-speed 1000 Hz camera (Photron, SA3) was positioned to observe one step, and the take-off and touchdown frames were manually annotated in this video. This dataset only contains trained athletes and running speeds were in the range 7 to 8.6 m/s.

The track had windows along both sides, as well as some artificial lighting. Weather was a changeable partly cloudy day, and the lighting varies dramatically from run to run, and also during some runs (examples can be seen in Fig. 19). The running lane was selected to be in shadow for the duration (note that just the feet are guaranteed to be in shadow) to avoid hard shadows from direct sunlight, but the overall ambient lighting and the brightness of the background, change substantially. The camera settings were not adjusted during the session, but lighting was generally good enough to allow the BGS to produce clean foreground masks.

This dataset consists of 70 runs of 14 trained athletes. Two of these athletes were hurdlers, and there was a single hurdle present in the observed area.

Fig. 17 Images showing the capture environments for the first (top) and second (bottom) datasets. The images are captured using one of the track-parallel cameras (see Fig. 1) allowing the running area and other cameras to be observed

Fig. 18 Camera system for the first dataset, showing the location of the 'a' and 'b' cameras (see Sect. 3) and the force plates

6.3 Comparison systems

The performance of the Obsmos and Fused OpenPose systems can be compared with results reported in the literature and with the system of Harle [23].

A variant of Harle's system was created and mildly adapted to the multi-camera environment. Foreground masks are accumulated over the course of each run. The resulting image shows how long each pixel of the image was labelled as foreground. This accumulator image can be thresholded to find regions that are static for a specified minimum time. The remaining regions in the thresholded map indicate the location of foot contacts. Harle then used various filtering tools to remove false positives, whereas the multicamera system used for Obsmos and Fused OpenPose allowed for simple cross-camera validation to achieve the same result. Timing of each event is taken by examining at what frame foreground started to accumulate, and stopped accumulating, as per the paper. Where contacts are visible in multiple cameras, the same latest-landing, earliest-take-off scheme used by Obsmos is used.

As seen in Table 5, Harle's approach suffers from the foreground noise of the first dataset leading to poor estimates of contact events. It does better in the second dataset where

Table 1 Landing, take-off and duration errors for contacts. μ is mean signed error, σ is the standard deviation of the signed error, and $\bar{\mu}$ is mean absolute error. All numbers are frames at 180 fps

algorithm		Obsmos			Fused OpenPose		
		Trained	Rec	Overall	Trained	Rec	Overall
landing	μ	−0.32	−0.78	−0.64	−0.02	−4.78	−2.87
	σ	0.75	2.02	1.75	0.81	7.74	6.45
	$\bar{\mu}$	0.61	1.35	1.13	0.61	5.59	3.99
take-off	μ	−0.78	−0.01	−0.24	0.87	−0.03	0.33
	σ	1.36	2.34	2.13	0.83	2.66	2.17
	$\bar{\mu}$	1.15	1.62	1.48	0.96	1.68	1.39
duration	μ	−0.46	0.77	0.40	−0.89	−4.75	−3.20
	σ	1.43	3.38	2.99	0.99	7.64	6.23
	$\bar{\mu}$	1.06	2.04	1.75	1.02	5.99	3.99

Table 2 Step length errors for the first dataset. Errors are given in mm. μ is mean signed error, σ is the standard deviation of the signed error, and $\bar{\mu}$ is mean absolute error

Algorithm	μ	σ	$\bar{\mu}$
Obsmos	1.2	7.37	6.11
Fused OpenPose	−0.88	16.02	12.01

Table 3 Step length errors for the second dataset. Errors are given in mm. μ is mean signed error, σ is the standard deviation of the signed error, and $\bar{\mu}$ is mean absolute error

system	μ	σ	$\bar{\mu}$
Obsmos	0.52	6.43	4.96
Fused OpenPose	1.68	24.66	19.14

the foreground masks are cleaner, but performance still lags behind that of the Obsmos system.

Step length results can be directly compared with the video system of Dunn [13] which was reported to have -4.9 ± 177.7 mm (Bland and Altman 95% limits of agreement). In comparison, the Obsmos system is more accurate (has less bias) and more precise, at 0.52 ± 12.58 mm for step length, and Fused OpenPose is also better at 1.68 ± 48.33 mm. Dunn reported contact time as -0.03 ± 0.03 s LOA against their annotations. Obsmos achieves 0.004 ± 0.008 s for landings on the second dataset, Fused OpenPose is -0.01 ± 0.02 s. In comparison with a system based on optical motion capture, Nagahara [38] reported contact event timings with a mean and standard deviation in the region of 0.5 ± 0.6 (absolute mean $\pm$ standard deviation) frames at 250 Hz, which is better than any reported video-based approach—however this is not surprising considering the quality of track for an optical marker.

Table 4 Step timing errors for the second dataset. Errors are given in frames at 180 fps. μ is mean signed error, σ is the standard deviation of the signed error, and $\bar{\mu}$ is mean absolute error

algorithm	Obsmos			Fused OpenPose		
	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
landing	0.71	0.75	0.91	−1.74	1.88	1.97
take-off	0.80	1.13	1.16	1.21	1.75	1.21
duration	0.08	0.98	0.84	1.42	1.43	1.57

Table 5 Step timing errors for the comparison system of Harle [23]. Errors are given in frames at 180 fps. μ is mean signed error, σ is the standard deviation of the signed error, and $\bar{\mu}$ is mean absolute error

Dataset	1. Laboratory			2. Track		
	μ	σ	$\bar{\mu}$	μ	σ	$\bar{\mu}$
landing	5.6	3.3	5.7	0.29	2.10	1.56
take-off	1.3	3.2	2.6	−0.93	1.56	1.31
duration	−4.3	3.5	4.6	−0.84	1.31	1.46

6.4 Discussion of results

On the first dataset, there is a clear division of performance between the recreational runners and the trained sprinters. For the trained sprinters, both Obsmos and Fuse OpenPose prove equally good on measuring landing events (absolute error of 0.61 frames at 180 fps), while the Fused OpenPose system showed a marginal advantage on take-off (0.96 versus 1.15 frames at 180 fps). For recreational runners, both systems showed very similar performance on take-off (1.62 and 1.68 frames at 180 fps), but Obsmos has a clear advantage on landing (1.35 versus 5.59 frames at 180 fps). This is not surprising however, as the recreational runners exhibit a wide variety of running styles, including heel-first or flat-foot landings. Where a runner has a heel-first style, the acceleration peak of the toe is not going to coincide with the moment of foot-ground contact. Where a recreational runner adopts a

Fig. 19 Strong lighting changes can be seen with significantly increased brightness through the windows creating strong highlights on the track and significant overall brightness changes to the background of cameras facing the window. These two images are taken from the same run. The running lane was selected so that the feet would not be exposed to the harsh direct sunlight, thus preventing difficult shadows around the feet, but the strong lighting changes still present a difficulty for background subtraction (athlete blurred to meet privacy requirements of the volunteers)

more toe-first sprinting style, the Fused OpenPose approach continued to perform well. If the aim is to solve for all variations in running style, a modification of the Fused OpenPose system would have to be made, possibly by including analysis of the heel or ankle joints as well as the toes. Obsmos does not suffer from this problem because it does not specifically prefer the toe or the heel.

The performance of the Obsmos system is consistent between the first and second datasets, whereas there is a notable degradation in performance for the Fused OpenPose system which adds, on average, a frame of absolute error. It is interesting to note however that when the runs are grouped for individual runners, the precision of Fused OpenPose remains high (typical per-runner standard deviations are in the range 0.5 to 0.9 frames) but the bias can vary substantially between individual runners, with one runner having a bias of 5 frames making the overall result less consistent. This suggests that the Fused OpenPose approach is more susceptible to different running technique and perhaps can only reach its best performance once the bias is known for individual runners.

For step length Obsmos is significantly better, with standard deviations of 7.37 and 6.43 mm versus 16.02 and 24.66 mm for the two datasets. This demonstrates the ability of Obsmos to consistently position the foot-box using all available camera views. The Fused OpenPose system does not optimise a toe position using multiple camera views, but it is at the mercy of the original OpenPose detections which although generally accurate can be imprecise, picking out slightly different parts of the foot depending on the exact camera angles. The result is that the measurement of step length is much less precise than for Obsmos and is also consistent with the positional accuracies reported by other 3D OpenPose papers which have suggested 3D joint position accuracies in the region of 20 to 40 mm [39]. Improving the step length accuracy of the Fused OpenPose approach could, in the worst case, mean that the dataset used to train these neural networks needs to be re-annotated by experts with far more care and consistency.

The largest weakness of the Obsmos system is the use of background subtraction. Where this is performed robustly,

which may require a degree of environmental control, excellent results can be expected. As background subtraction is a well-researched and widely used solution, its limitations are well known and often controllable. Alternatives do exist in the form of human segmentation algorithms [20,33], but these have a very significant speed trade-off (seconds per frame instead of tens of frames per second) and might not be adjustable if they happen to not work in some environment. As it is, a modest level of environmental control and the use of a robust and easily tunable BGS implementation such as IMBS-MT [6] can be highly effective.

7 Conclusion

Two approaches to measuring the timing and location of a sprinter's foot-ground contacts have been presented and compared. The first approach—Obsmos—is an occupancy map-based approach building off the previous work of [15], and the second approach—Fused OpenPose—fuses monocular machine learning-based pose prediction across multiple cameras and then extracts foot contact events from tracks of the detected toes. Machine learning systems have shown the ability to greatly improve the performance of various computer vision tasks, but they require large training datasets that can be impractical to create for niche, specialist applications that require high accuracy. The comparison explores whether a general purpose machine learning-based system can be readily adapted to a specialist application, and how it compares to a traditionally engineered algorithm specific to the task.

An evaluation of the two approaches has been presented using two datasets, one in a laboratory environment and the other in a more real-world scenario. Both systems had good and similar performance for estimating contact timing, with the Obsmos system being slightly more consistent between datasets. Obsmos was also significantly more precise for measuring step lengths.

A number of trade-offs were encountered which could inform future adoption of either technique. The Obsmos system depends upon many processing stages which might need to be tuned for the specific environment. One particular weakness is the use of background subtraction for human segmentation which has many known limitations (harsh lighting, busy and changeable backgrounds, strong shadows, poor image contrast, etc...). A machine learning-based human segmentation could in future be used to replace the background subtraction stage but at substantial cost in processing time. The Fused OpenPose system needed no specific tuning for the environment, but there is a suggestion that optimal performance on foot-ground contact timing might need tuning to an individual runner, and imprecision of toe localisation led to relatively poor estimates of step length.

Acknowledgements This research was funded by CAMERA, the RCUK Centre for the Analysis of Motion, Entertainment Research and Applications, EP/M023281/1.

Author Contributions All authors contributed to the study conception and design. Data collection and analysis were performed by Murray Evans and Steffi Colyer. Athletes were recruited by Steffi Colyer and Aki Salo. Software was developed by Murray Evans. Analysis was performed by Murray Evans and Steffi Colyer. Supervision was provided by Aki Salo and Darren Cosker. Funding was acquired by Darren Cosker and Aki Salo. Writing—original drafts prepared by Murray Evans with review by Darren Cosker. All authors provided further review and approval.

Funding This research was funded by CAMERA, the RCUK Centre for the Analysis of Motion, Entertainment Research and Applications, EP/M023281/1.

Availability of data The video data collected and used for this study cannot be shared due to agreements with the volunteers.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Software used for this study is not being made available.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Allison, T.: More than a man in a monkey suit: andy serkis, motion capture, and digital realism. *Q Rev Film Video* **28**(4), 325–341 (2011). <https://doi.org/10.1080/10509208.2010.500947>
2. Amini, A., Banitsas, K., Hosseinzadeh, S.: A new technique for foot-off and foot contact detection in a gait cycle based on the knee joint angle using microsoft kinect v2. In: 2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pp. 153–156 (2017). <https://doi.org/10.1109/BHI.2017.7897228>
3. Bezodis, I., Salo, A.I.T., Kerwin, D.: A longitudinal case study of step characteristics in a world class sprint athlete. In: Proceedings of XXVI International Conference on Biomechanics in Sports, pp. 537–540 (2008)
4. Bezodis, I., Thomson, A., Gittoes, M., Kerwin, D.: Identification of instants of touchdown and take-off in sprint running using an automatic motion analysis system. In: 25th International Symposium on Biomechanics in sports, pp. 501–504 (2007)

5. Bezodis, N.E., Salo, A.I.T., Trewartha, G.: Relationships between lower-limb kinematics and block phase performance in a cross section of sprinters. *Eur. J. Sport Sci.* **15**(2), 118–124 (2015)
6. Bloisi, D.D., Pennisi, A., Iocchi, L.: Parallel multi-modal background modeling. *Pattern Recognit. Lett.* **96**(Suppl C), 45–54 (2017). <https://doi.org/10.1016/j.patrec.2016.10.016>
7. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: An overview. *Comput. Sci. Rev.* **11**(Suppl C), 31–66 (2014). <https://doi.org/10.1016/j.cosrev.2014.04.001>
8. Caldas, R., Mundt, M., Potthast, W., Lima Neto, F., Markert, B.: A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait & Posture* **57**, 204–210 (2017). <https://doi.org/10.1016/j.gaitpost.2017.06.019>
9. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint [arXiv:1812.08008](https://arxiv.org/abs/1812.08008) (2018)
10. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
11. Colyer, S., Evans, M., Cosker, D., Salo, A.: A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Med. Open* (2018). <https://doi.org/10.1186/s40798-018-0139-y>
12. Devita, P., Hortobagyi, T., Barrier, J., et al.: Gait biomechanics are not normal after anterior cruciate ligament reconstruction and accelerated rehabilitation. *Med. Sci. Sports Exerc.* **30**, 1481–1488 (1998)
13. Dunn, M., Kelley, J.: Non-invasive, spatio-temporal gait analysis for sprint running using a single camera. *Proc. Eng.* **112**, 528–533 (2015). <https://doi.org/10.1016/j.proeng.2015.07.237>
14. El-Sallam, A., Bennamoun, M., Honda, K., Lyttle, A., Alderson, J.: Towards a fully automatic markerless motion analysis system for the estimation of body joint kinematics with application to sport analysis. In: 10th International Conference on Computer Graphics Theory and Applications (VISIGRAPP), pp. 58–69 (2015)
15. Evans, M., Colyer, S.L., Cosker, D., Salo, A.I.T.: Foot contact timings and step length for sprint training. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1652–1660 (2018)
16. Fabbri, R., Costa, L.D.F., Torelli, J.C., Bruno, O.M.: 2d euclidean distance transform algorithms: A comparative survey. *ACM Comput. Surv.* **40**(1), 2:1–2:44 (2008). <https://doi.org/10.1145/1322432.1322434>
17. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), 381–395 (1981). <https://doi.org/10.1145/358669.358692>
18. Ford, K.R., Myer, G.D., Toms, H.E., Hewett, T.E.: Gender differences in the kinematics of unanticipated cutting in young athletes. *Med. Sci. Sports Exerc.* **37**(1), 124–129 (2005)
19. Frost, D., Cronin, J.: Stepping back to improve sprint performance: a kinetic analysis of the first step forwards. *J. Strength Cond. Res.* **25**, 2721–2728 (2011)
20. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. arXiv (2018)
21. Halvorsen, P., Sægrov, S., Mortensen, A., Kristensen, D., Eichhorn, A., Stenhaug, M., Dahl, S., Stensland, H., Gaddam, V., Griwodz, C., Johansen, D.: Bagadus: An integrated system for arena sports analytics - a soccer case study. In: Proceedings of the 4th ACM Multimedia Systems Conference, MMSys 2013, pp. 48–59 (2013). <https://doi.org/10.1145/2483977.2483982>
22. Handsaker, J.C., Forrester, S.E., Folland, J.P., Black, M.I., Allen, S.J.: A kinematic algorithm to identify gait events during running at different speeds and with different footstrike types. *J. Biomech.* **49**(16), 4128–4133 (2016). <https://doi.org/10.1016/j.jbiomech.2016.10.013>
23. Harle, R., Cameron, J., Lasenby, J.: Foot contact detection for sprint training. In: Asian Conference on Computer Vision 2010 (ACCV2010), ACCV'10, pp. 297–306. Springer-Verlag, Berlin, Heidelberg (2010). <http://dl.acm.org/citation.cfm?id=2040690.2040722>
24. Harris, G., Wertsch, J.: Procedure for gait analysis. *Arch. Phys. Med. Rehabil.* **75**, 216–225 (1994)
25. Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B.: Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. *IEEE J. Sel. Topics Signal Process.* **6**(5), 538–552 (2012)
26. Jarchi, D., Pope, J., Lee, T.K.M., Tamjidi, L., Mirzaei, A., Sanei, S.: A review on accelerometry-based gait analysis and emerging clinical applications. *IEEE Rev. Biomed. Eng.* **11**, 177–194 (2018). <https://doi.org/10.1109/RBME.2018.2807182>
27. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. *Proc. Br. Mach. Vis. Conf.* (2010). <https://doi.org/10.5244/C.24.12>
28. Johnson, S.G.: The nlopt nonlinear-optimization package. <http://ab-initio.mit.edu/nlopt> (2018)
29. Jung, S., Nixon, M.S.: Heel strike detection based on human walking movement for surveillance analysis. *Pattern Recogn. Lett.* **34**(8), 895–902 (2013). <https://doi.org/10.1016/j.patrec.2012.08.007>
30. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
31. Khan, S., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(3), 505–519 (2009). <https://doi.org/10.1109/TPAMI.2008.102>
32. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
33. Lin, K., Wang, L., Luo, K., Chen, Y., Liu, Z., Sun, M.T.: Cross-domain complementary learning with synthetic data for multi-person part segmentation. arXiv preprint [arXiv:1907.05193](https://arxiv.org/abs/1907.05193) (2019)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014, pp. 740–755. Springer, Cham (2014)
35. Magnor, M.A., Grau, O., Sorkine-Hornung, O., Theobalt, C.: Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality. AK Peters/CRC Press, Boca Raton (2015)
36. Maiwald, C., Sterzing, T., Mayer, T., Milani, T.: Detecting foot-to-ground contact from kinematic data in running. *Footwear Sci.* **1**(2), 111–118 (2009)
37. Milner, C.E., Paquette, M.R.: A kinematic method to detect foot contact during running for all foot strike patterns. *J. Biomech.* **48**(12), 3502–3505 (2015). <https://doi.org/10.1016/j.jbiomech.2015.07.036>
38. Nagahara, R., Zushi, K.: Determination of foot strike and toe-off event timing during maximal sprint using kinematic data. *Int. J. Sport Health Sci.* **11**, 96–100 (2013)
39. Nakano, N., Sakura, T., Ueda, K., Omura, L., Kimura, A., Iino, Y., Fukushima, S., Yoshioka, S.: Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. *Front. Sports Active Living* **2**, 50 (2020). <https://doi.org/10.3389/fspor.2020.00050>
40. Nithyakani, P. et al. The systematic review on gait analysis: trends and developments. *Eur. J. Mol. Clin. Med.* **7**(6), 1636–1654 (2020)

41. Osis, S., Hettinga, B., Leitch, J., Ferber, R.: Predicting timing of foot strike during running, independent of striking technique, using principal component analysis of joint angles. *J. Biomech.* (2014). <https://doi.org/10.1016/j.jbiomech.2014.06.009>
42. Ringer, M., Lasenby, J.: A procedure for automatically estimating model parameters in optical motion capture. *Image Vis. Comput.* **22**(10), 843–850 (2004)
43. Seethapathi, N., Wang, S., Saluja, R., Blohm, G., Kording, K.P.: Movement science needs different pose tracking algorithms (2019)
44. Silaghi, M.C., Plänkers, R., Boulic, R., Fua, P., Thalmann, D.: Local and global skeleton fitting techniques for optical motion capture. In: *International Workshop on Capture Techniques for Virtual Environments*, pp. 26–40. Springer (1998)
45. Sjö Dahl, C., Jarnlo, G.B., Söderberg, B., Persson, B.: Kinematic and kinetic gait analysis in the sagittal plane of trans-femoral amputees before and after special gait re-education. *Prosthet. Orthot. Int.* **26**(2), 101–112 (2002)
46. Slawinski, J., Bonnefoy, A., Levêque, J.M., Ontanon, G., Riquet, A., Dumas, R., Chêze, L.: Kinematic and kinetic comparisons of elite and well-trained sprinters during sprint start. *J. Strength Cond. Res.* **24**(4), 896–905 (2010)
47. Starke, S., Clayton, H.: A universal approach to determine footfall timings from kinematics of a single foot marker in hoofed animals. *PeerJ* (2015). <https://doi.org/10.7717/peerj.783>
48. Tompson, J., Goroshin, R., Jain, A., Lecun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 648–656 (2015). <https://doi.org/10.1109/CVPR.2015.7298664>
49. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660 (2014). <https://doi.org/10.1109/CVPR.2014.214>
50. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pp. 298–372. Springer Berlin Heidelberg, Berlin, Heidelberg (2000). https://doi.org/10.1007/3-540-44480-7_21. https://doi.org/10.1007/3-540-44480-7_21
51. Windolf, M., Götz, N., Morlock, M.: Systematic accuracy and precision analysis of video motion capturing systems-exemplified on the vicon-460 system. *J. Biomech.* **41**(12), 2776–2780 (2008). <https://doi.org/10.1016/j.jbiomech.2008.06.024>
52. Xu, X., McGorry, R., Chou, L.S., Lin, J.H., Chang, C.C.: Accuracy of the microsoft kinecttm for measuring gait parameters during treadmill walking. *Gait Posture* (2015). <https://doi.org/10.1016/j.gaitpost.2015.05.002>
53. Yang, S.X., Christiansen, M.S., Larsen, P.K., Alkjær, T., Moeslund, T.B., Simonsen, E.B., Lynnerup, N.: Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2**(1), 46–65 (2014)
54. Zago, M., Luzzago, M., Marangoni, T., Cecco, M., Tarabini, M., Galli, M.: 3d tracking of human motion using visual skeletonization and stereoscopic vision. *Front. Bioeng. Biotechnol.* (2020). <https://doi.org/10.3389/fbioe.2020.00181>
55. Zhang, C., Yang, F., Li, G., Zhai, Q., Jiang, Y., Xuan, D.: Mv-sports: a motion and vision sensor integration-based sports analysis system. *IEEE INFOCOM 2018—IEEE Conference on Computer Communications* pp. 1070–1078 (2018)
56. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334 (2000). <https://doi.org/10.1109/34.888718>
57. Zhu, W., Anderson, B., Zhu, S., Wang, Y.: A computer vision-based system for stride length estimation using a mobile phone camera. In: *Proceedings of the 18th International ACM SIGAC-*

CESS Conference on Computers and Accessibility, ASSETS '16, pp. 121–130. ACM, New York, NY, USA (2016). <https://doi.org/10.1145/2982142.2982156>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Murray Evans is a Research Associate with the EPSRC-funded Centre for the Analysis of Motion Entertainment Research and Applications (CAMERA) at the University of Bath, UK. He completed his PhD in stereo computer vision at the University of Reading, UK, and continued in a research role making contributions and taking technical lead on various EU-funded projects related to video surveillance (AVITRACK, ISCAPS, SAFEE, CoFRIEND, EFFISEC, ARENA) as well as significant contributions to the well-known PETS datasets. He spent two years as a research scientist at Sharp Labs of Europe before joining CAMERA to work on cross-domain research projects applying computer vision technologies to sports and markerless motion capture. Murray's research interests include many aspects of computer vision, but especially its application in novel or niche domains.

Steffi Colyer is a Lecturer in Biomechanics at the University of Bath, UK. She received her B.Sc. degree in Sport and Exercise Science in 2011 and Ph.D. with the Department for Health at the University of Bath in 2015. She worked as a Research Associate within the EPSRC-funded Centre for the Analysis of Motion, Entertainment and Research Applications (CAMERA) at the University of Bath for four years before becoming a Lecturer in Biomechanics in 2019. Her research interests include sports biomechanics, the application of markerless technology, and musculoskeletal and biomechanical adaptations to training.

Dr Aki Salo is currently the Director of the Research Institute for Olympic Sports, Finland, and honorary reader at the University of Bath, UK. He is a Chartered Scientist as well as a Fellow of the British Association of Sport and Exercise Sciences and the International Society of Biomechanics in Sports. After obtaining his PhD from the University of Exeter in 1999, UK, Aki worked 20 years at the University of Bath. He has published over 70 scientific articles mainly in the area of Sports Science.

Darren Cosker is Professor in Computer Science at the University of Bath and Director of the Centre for the Analysis of Motion, Entertainment Research and Applications (CAMERA). He has been awarded research fellowships from the Royal Academy of Engineering (2007–2012) and Royal Society Industry Fellowship (2012–2016). Although primarily embedded in visual computing and AI, he is interested in problems that cut across disciplines and has published widely including such areas as animation, vision, emotion perception and sport.